Towards Practical Planning to Predict and Exploit Intentions for Interacting with Self-Interested Agents

Trong Nghia Hoang & Kian Hsiang Low
Department of Computer Science, National University of Singapore
Republic of Singapore
{nghiaht, lowkh}@comp.nus.edu.sg

Abstract

A key challenge in non-cooperative multi-agent systems is that of developing efficient planning algorithms for intelligent agents to interact and perform effectively among boundedly rational, self-interested agents (e.g., humans). The practicality of existing works addressing this challenge is being undermined due to either the restrictive assumptions of the other agents' behavior, the failure in accounting for their rationality, or the prohibitively expensive cost of modeling and predicting their intentions. To boost the practicality of research in this field, we investigate how intention prediction can be efficiently exploited and made practical in planning, thereby leading to efficient intention-aware planning frameworks capable of predicting the intentions of other agents and acting optimally with respect to their predicted intentions. We show that the performance losses incurred by the resulting planning policies are linearly bounded by the error of intention prediction. Empirical evaluations through a series of stochastic games demonstrate that our policies can achieve better and more robust performance than the state-of-the-art algorithms.

1 Introduction

A fundamental challenge in non-cooperative *multi-agent systems* (MAS) is that of designing intelligent agents that can efficiently plan their actions under uncertainty to interact and perform effectively among boundedly rational[1], self-interested agents (e.g., humans). Such a challenge is posed by many real-world applications [7], which include automated electronic trading markets where software agents interact, and traffic intersections where autonomous cars have to negotiate with human-driven vehicles to cross them, among others. These applications can be modeled as partially observable stochastic games (POSGs) in which the agents are self-interested (i.e., non-cooperative) and do not necessarily share the same goal, thus invalidating the use of planning algorithms developed for coordinating cooperative agents (i.e., solving POSGs with common payoffs) [11, 15, 16]. Existing planning frameworks for non-cooperative MAS can be generally classified into:

Game-theoretic frameworks. Based on the well-founded classical game theory, these multi-agent planning frameworks [7, 9][2] characterize the agents' interactions in a POSG using solution concepts such as Nash equilibrium. Such frameworks suffer from the following drawbacks: (a) Multiple equilibria may exist, (b) only the optimal actions corresponding to the equilibria are specified, and (c) they assume that the agents do not collaborate to beneficially deviate from the equilibrium (i.e., no coalition), which is often violated by human agents.

[1] Boundedly rational agents are subject to limited cognition and time in making decisions [5].

[2] The learning framework of Hu and Wellman [1998] trivially reduces to planning when the transition model is known a priori.

Decision-theoretic frameworks. Unafflicted by the drawbacks of game-theoretic approaches, they extend single-agent decision-theoretic planning frameworks such as the *Markov decision process* (MDP) and *partially observable Markov decision process* (POMDP) to further characterize interactions with the other self-interested agents in a POSG. In particular, the *interactive POMDP* (I-POMDP) framework [3, 4, 6, 14] is proposed to explicitly account for the bounded rationality of self-interested agents: It replaces POMDP's flat beliefs over the physical states with interactive beliefs over both the physical states and the other agent's beliefs. Empowered by such an enriched, highly expressive belief space, I-POMDP can explicitly model and predict the other agent's intention (i.e., mixed strategy) under partial observability.

However, solving I-POMDP is prohibitively expensive due to the following computational difficulties [4, 6]: (a) **Curse of dimensionality** – since I-POMDP's interactive belief is over the joint space of physical states and the other agent's beliefs (termed interactive state space in [4]), its dimension can be extremely large and possibly infinite; (b) **curse of history** – similar to POMDP, I-POMDP's policy space grows exponentially with the length of planning horizon; and (c) **curse of nested reasoning** – as I-POMDP utilizes a nested structure within our agent's belief space to represent its belief over the other agent's belief and the other agent's belief over our agent's belief and so on, it aggravates the effects of the other two curses [4].

To date, a number of approximate I-POMDP techniques [3, 4, 14] have been proposed to mitigate some of the above difficulties. Notably, *Interactive Particle Filtering* (I-PF) [3] focused on alleviating the curse of dimensionality by generalizing the particle filtering technique to accommodate the multi-agent setting while *Interactive Point-based Value Iteration* [4] (I-PBVI) aimed at relieving the curse of history by generalizing the well-known point-based value iteration (PBVI) [13] to operate in the interactive belief space. Unfortunately, I-PF fails to address the curse of history and it is not clear how PBVI or other sampling-based algorithm can be modified to work with a particle representation of interactive beliefs, whereas I-PBVI suffers from the curse of dimensionality because its dimension of interactive belief grows exponentially with the length of planning horizon of the other agent (Section 3). Using interactive beliefs, it is therefore not known whether it is even possible to jointly lift both curses, for example, by extending I-PF or I-PBVI. Furthermore, they do not explicitly account for the curse of nested reasoning. As a result, their use has been restricted to small, simple problems [12] (e.g., multiagent Tiger [6, 11, 4]).

To tractably solve larger problems, existing approximate I-POMDP techniques such as I-PF and I-PBVI have to significantly reduce the quality of approximation and impose restrictive assumptions (Section 3), or risk not producing a policy at all with the available memory of modern-day computers. This naturally raises the concern of whether the resulting policy can still perform well or not under different partially observable environments, as investigated in Section 5. Since such drastic compromises in solution quality are necessary of approximate I-POMDP techniques to tackle a larger problem directly, it may be worthwhile to instead consider formulating an approximate version of the problem with a less sophisticated structural representation such that it allows an exact or near-optimal solution policy to be more efficiently derived. More importantly, can the induced policy perform robustly against errors in modeling and predicting the other agent's intention? If we are able to formulate such an approximate problem, the resulting policy can potentially perform better than an approximate I-POMDP policy in the original problem while incurring significantly less planning time.

Our work in this paper investigates such an alternative: We first develop a novel intention-aware *nested MDP* framework (Section 2) for planning in fully observable multi-agent environments. Inspired by the cognitive hierarchy model of games [1], nested MDP constitutes a recursive reasoning formalism to predict the other agent's intention and then exploit it to plan our agent's optimal interaction policy. Its formalism is by no means a reduction of I-POMDP. We show that nested MDP incurs linear time in the planning horizon length and reasoning depth. Then, we propose an I-POMDP Lite framework (Section 3) for planning in partially observable multi-agent environments that, in particular, exploits a practical structural assumption: The intention of the other agent is driven by nested MDP, which is demonstrated theoretically to be an effective surrogate of its true intention when the agents have fine sensing and actuation capabilities. This assumption allows the other agent's intention to be predicted efficiently and, consequently, I-POMDP Lite to be solved efficiently in polynomial time, hence lifting the three curses of I-POMDP. As demonstrated empirically, it also improves I-POMDP Lite's robustness in planning performance by overestimating the true sensing capability of the other agent. We provide theoretical performance guarantees of the nested MDP and I-POMDP Lite policies that improve with decreasing error of intention prediction (Section 4). We extensively evaluate our frameworks through experiments involving a series of POSGs that have to be modeled using a significantly larger state space (Section 5).

2 Nested MDP

Given that the environment is fully observable, our proposed nested MDP framework can be used to predict the other agent's strategy and such predictive information is then exploited to plan our agent's optimal interaction policy. Inspired by the cognitive hierarchy model of games [1], it constitutes a well-defined recursive reasoning process that comprises k levels of reasoning. At level 0 of reasoning, our agent simply believes that the other agent chooses actions randomly and computes its best response by solving a conventional MDP that implicitly represents the other agent's actions as stochastic noise in its transition model. At higher reasoning levels $k \geq 1$, our agent plans its optimal strategy by assuming that the other agent's strategy is based only on lower levels $0, 1, \ldots, k-1$ of reasoning. In this section, we will formalize nested MDP and show that our agent's optimal policy at level k can be computed recursively.

Nested MDP Formulation. Formally, nested MDP for agent t at level k of reasoning is defined as a tuple $M^k_t \triangleq \big(S, U, V, T, R, \{\pi^i_{-t}\}^{k-1}_{i=0}, \phi\big)$ where S is a set of all possible states of the environment; U and V are, respectively, sets of all possible actions available to agents t and -t; $T : S \times U \times V \times S \to [0, 1]$ denotes the probability $Pr(s'|s, u, v)$ of going from state $s \in S$ to state $s' \in S$ using agent t's action $u \in U$ and agent -t's action $v \in V$; $R : S \times U \times V \to \mathbb{R}$ is a reward function of agent t; $\pi^i_{-t} : S \times V \to [0, 1]$ is a reasoning model of agent -t at level $i < k$, as defined later in (3); and $\phi \in (0, 1)$ is a discount factor.

Nested MDP Planning. The optimal $(h + 1)$-step-to-go value function of nested MDP M^k_t at level $k \geq 0$ for agent t satisfies the following Bellman equation:

$$\begin{aligned} U^{k,h+1}_t(s) &\triangleq \max_{u \in U} \sum_{v \in V} \widehat{\pi}^k_{-t}(s, v)\; Q^{k,h+1}_t(s, u, v) \\ Q^{k,h+1}_t(s, u, v) &\triangleq R(s, u, v) + \phi \sum_{s' \in S} T(s, u, v, s')\; U^{k,h}_t(s') \end{aligned} \tag{1}$$

where the mixed strategy $\widehat{\pi}^k_{-t}$ of the other agent -t for $k > 0$ is predicted as

$$\widehat{\pi}^k_{-t}(s, v) \triangleq \begin{cases} \sum^{k-1}_{i=0} p(i)\pi^i_{-t}(s, v) & \text{if } k > 0, \\ |V|^{-1} & \text{otherwise.} \end{cases} \tag{2}$$

where the probability $p(i)$ (i.e., $\sum^{k-1}_{i=0} p(i) = 1$) specifies how likely agent -t will reason at level i; a uniform distribution is assumed when there is no such prior knowledge. Alternatively, one possible direction for future work is to learn $p(i)$ using multi-agent reinforcement learning techniques such as those described in [2, 8]. At level 0, agent -t's reasoning model π^0_{-t} is induced by solving M^0_{-t}. To obtain agent -t's reasoning models $\{\pi^i_{-t}\}^{k-1}_{i=1}$ at levels $i = 1, \ldots, k-1$, let $Opt^i_{-t}(s)$ be the set of agent -t's optimal actions for state s induced by solving its nested MDP M^i_{-t}, which recursively involves building agent t's reasoning models $\{\pi^l_t\}^{i-1}_{l=0}$ at levels $l = 0, 1, \ldots, i-1$, by definition. Then,

$$\pi^i_{-t}(s, v) \triangleq \begin{cases} |Opt^i_{-t}(s)|^{-1} & \text{if } v \in Opt^i_{-t}(s), \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

After predicting agent -t's mixed strategy $\widehat{\pi}^k_{-t}$ (2), agent t's optimal policy (i.e., reasoning model) π^k_t at level k can be induced by solving its corresponding nested MDP M^k_t (1).

Time Complexity. Solving M^k_t involves solving $\{M^i_{-t}\}^{k-1}_{i=0}$, which, in turn, requires solving $\{M^i_t\}^{k-2}_{i=0}$, and so on. Thus, solving M^k_t requires solving M^i_t $(i = 0, \ldots, k-2)$ and M^i_{-t} $(i = 0, \ldots, k-1)$, that is, $\mathcal{O}(k)$ nested MDPs. Given $\widehat{\pi}^k_{-t}$, the cost of deriving agent t's optimal policy grows linearly with the horizon length h as the backup operation (1) has to be performed h times. In turn, each backup operation incurs $\mathcal{O}\big(|S|^2\big)$ time given that $|U|$ and $|V|$ are constants. Then, given agent -t's profile of reasoning models $\{\pi^i_{-t}\}^{k-1}_{i=0}$, predicting its mixed strategy $\widehat{\pi}^k_{-t}(s, v)$

(2) incurs $\mathcal{O}(k)$ time. Therefore, solving agent t's nested MDP M_t^k (1) or inducing its corresponding reasoning model π_t^k incurs $\mathcal{O}\big(kh|S|^2\big)$.

3 Intention-Aware POMDP

To tackle partial observability, it seems obvious to first consider generalizing the recursive reasoning formalism of nested MDP. This approach yields two practical complications: (a) our agent's belief over both the physical states and the other agent's beliefs (i.e., a probability distribution over probability distributions) has to be modeled, and (b) the other agent's mixed strategy has to be predicted for each of its infinitely many possible beliefs. Existing approximate I-POMDP techniques address these respective difficulties by (a) using a finite particle representation like I-PF [3] or (b) constraining the interactive state space IS to $IS' = S \times \text{Reach}(B, h)$ like I-PBVI [4] where $\text{Reach}(B, h)$ includes the other agent's beliefs reachable from a finite set B of its candidate initial beliefs over horizon length h.

However, recall from Section 1 that since I-PF suffers from the curse of history, the particle approximation of interactive beliefs has to be made significantly coarse to solve larger problems tractably, thus degrading its planning performance. I-PBVI, on the other hand, is plagued by the curse of dimensionality due to the need of constructing the set $\text{Reach}(B, h)$ whose size grows exponentially with h. As a result, it cannot tractably plan beyond a few look-ahead steps for even the small test problems in Section 5. Furthermore, it imposes a restrictive assumption that the true initial belief of the other agent, which is often not known in practice, needs to be included in B to satisfy the absolute continuity condition of interactive beliefs [4] (see Appendix C for more details). So, I-PBVI may not perform well under practical environmental settings where a long planning horizon is desirable or the other agent's initial belief is not included in B. For I-PF and I-PBVI, the curse of nested reasoning aggravates the effects of other curses.

Since predicting the other agent's intention using approximate I-POMDP techniques is prohibitively expensive, it is practical to consider a computationally cheaper yet credible information source providing its intention such as its nested MDP policy. Intuitively, such a policy describes the intention of the other agent with full observability who believes that our agent has full observability as well. Knowing the other agent's nested MDP policy is especially useful when the agents' sensing and actuation capabilities are expected to be good (i.e., accurate observation and transition models), as demonstrated in the following simple result:

Theorem 1. *Let $\widehat{Q}^n_{\text{-}t}(s, v) \triangleq |U|^{-1} \sum_{u \in U} Q^{0,n}_{\text{-}t}(s, v, u)$ and $\widehat{Q}^n_{\text{-}t}(b, v)$ denote n-step-to-go values of selecting action $v \in V$ in state $s \in S$ and belief b, respectively, for the other agent -t using nested MDP and I-POMDP at reasoning level 0 (i.e., MDP and POMDP). If $b(s) \geq 1 - \epsilon$ and*

$$\forall\,(s, v, u)\,\exists\,(s', o)\; Pr(s'|s, v, u) \geq 1 - \frac{\epsilon}{2} \;\wedge\; Pr(o|s, v) \geq 1 - \frac{\epsilon}{2}$$

for some $\epsilon \geq 0$, then

$$\left|\widehat{Q}^n_{\text{-}t}(s, v) - \widehat{Q}^n_{\text{-}t}(b, v)\right| \leq \epsilon\, \mathcal{O}\!\left(\frac{R_{\max} - R_{\min}}{1 - \phi}\right) \tag{4}$$

where $R_{\max}$ and $R_{\min}$ denote agent -t's maximum and minimum immediate payoffs, respectively.

Its proof is given in Appendix A. Following from Theorem 1, we conjecture that, as ϵ decreases (i.e., observation and transition models become more accurate), the nested MDP policy $\pi^0_{\text{-}t}$ of the other agent is more likely to approximate the exact I-POMDP policy closely. Hence, the nested MDP policy serves as an effective surrogate of the exact I-POMDP policy (i.e., true intention) of the other agent if the agents have fine sensing and actuation capabilities; such a condition often holds for typical real-world environments.

Motivated by the above conjecture and Theorem 1, we propose an alternative I-POMDP Lite framework by exploiting the following structural assumption: The intention of the other agent is driven by nested MDP. This assumption allows the other agent's intention to be predicted efficiently by computing its nested MDP policy, thus lifting I-POMDP's curse of nested reasoning (Section 2). More importantly, it enables both the curses of dimensionality and history to be lifted, which makes solving I-POMDP Lite very efficient, as explained below. Compared to existing game-theoretic frameworks [9, 10] which make strong assumptions of the other agent's behavior, our assumption is clearly less restrictive. Unlike the approximate I-POMDP techniques, it does not cause I-POMDP Lite to be subject to coarse approximation when solving larger problems, which can potentially result in better planning performance. Furthermore, by modeling and predicting the other agent's intention using nested MDP, I-POMDP Lite tends to overestimate its true sensing capability and can therefore achieve a more robust performance than I-PBVI using significantly less planning time under different partially observable environments (Section 5).

I-POMDP Lite Formulation. Our I-POMDP Lite framework constitutes an integration of the nested MDP for predicting the other agent's mixed strategy into a POMDP for tracking our agent's belief in partially observable environments. Naively, this can be achieved by extending the belief space to $\Delta(S \times V)$ (i.e., each belief b is now a probability distribution over the state-action space $S \times V$) and solving the resulting augmented POMDP. The size of representing each belief therefore becomes $\mathcal{O}(|S||V|)$ (instead of $\mathcal{O}(|S|)$), which consequently increases the cost of processing each belief (i.e., belief update). Fortunately, our I-POMDP Lite framework can alleviate this extra cost: By factorizing $b(s,v) = b(s)\ \widehat{\pi}^{k}_{\text{-}t}(s,v)$, the belief space over $S \times V$ can be reduced to one over S because the predictive probabilities $\widehat{\pi}^{k}_{\text{-}t}(s,v)$ (2) (i.e., predicted mixed strategy of the other agent) are derived separately in advance by solving nested MDPs. This consequently alleviates the curse of dimensionality pertaining to the use of interactive beliefs, as discussed in Section 1. Furthermore, such a reduction of the belief space decreases the time and space complexities and typically allows an optimal policy to be derived faster in practice: the space required to store n sampled beliefs is only $\mathcal{O}(n|S| + |S||V|)$ instead of $\mathcal{O}(n|S||V|)$.

Formally, I-POMDP Lite (for our agent t) is defined as a tuple $(S, U, V, O, T, Z, R, \widehat{\pi}^{k}_{\text{-}t}, \phi, b_0)$ where S is a set of all possible states of the environment; U and V are sets of all actions available to our agent t and the other agent -t, respectively; O is a set of all possible observations of our agent t; $T : S \times U \times V \times S \to [0,1]$ is a transition function that depends on the agents' joint actions; $Z : S \times U \times O \to [0,1]$ denotes the probability $Pr(o|s', u)$ of making observation $o \in O$ in state $s' \in S$ using our agent t's action $u \in U$; $R : S \times U \times V \to \mathbb{R}$ is the reward function of agent t; $\widehat{\pi}^{k}_{\text{-}t} : S \times V \to [0,1]$ denotes the predictive probability $Pr(v|s)$ of selecting action v in state s for the other agent -t and is derived using (2) by solving its nested MDPs at levels $0, \ldots, k-1$; $\phi \in (0,1)$ is a discount factor; and $b_0 \in \Delta(S)$ is a prior belief over the states of environment.

I-POMDP Lite Planning. Similar to solving POMDP (except for a few modifications), the optimal value function of I-POMDP Lite for our agent t satisfies the below Bellman equation:

$$V_{n+1}(b) = \max_{u} \Big(R(b,u) + \phi \sum_{v,o} Pr(v,o|b,u)\ V_n(b') \Big) \tag{5}$$

where our agent t's expected immediate payoff is

$$R(b,u) = \sum_{s,v} R(s,u,v)\ Pr(v|s)\ b(s) \tag{6}$$

and the belief update is given as

$$b'(s') = \beta\ Z(s',u,o) \sum_{s} T(s,u,v,s')\ Pr(v|s)\ b(s)\ .$$

Note that (6) yields an intuitive interpretation: The uncertainty over the state of the environment can be factored out of the prediction of the other agent -t's strategy by assuming that agent -t can fully observe the environment. Consequently, solving I-POMDP Lite (5) involves choosing the policy that maximizes the expected total reward with respect to the prediction of agent -t's mixed strategy using nested MDP. Like POMDP, the optimal value function $V_n(b)$ of I-POMDP Lite can be approximated arbitrarily closely (for infinite horizon) by a piecewise-linear and convex function that takes the form of a set V_n[3] of α vectors:

$$V_n(b) = \max_{\alpha \in V_n} (\alpha \cdot b)\ . \tag{7}$$

Solving I-POMDP Lite therefore involves computing the corresponding set of α vectors that can be achieved inductively: given a finite set V_n of α vectors, we can plug (7) into (5) to derive V_{n+1} (see Theorem 3 in Section 4). Similar to POMDP, the number of α vectors grows exponentially

[3] With slight abuse of notation, the value function is also used to denote the set of corresponding α vectors.

with the time horizon: $|V_{n+1}| = |U||V_n|^{|V||O|}$. To avoid this exponential blow-up, I-POMDP Lite inherits essential properties from POMDP (Section 4) that make it amenable to be solved by existing sampling-based algorithm such as PBVI [13] used here. The idea is to sample a finite set B of reachable beliefs (from b_0) to approximately represent the belief simplex, thus avoiding the need to generate the full belief reachability tree to compute the optimal policy. This alleviates the curse of history pertaining to the use of interactive beliefs (Section 1). Then, it suffices to maintain a single α vector for each belief point $b \in B$ that maximizes $V_n(b)$. Consequently, each backup step can be performed in polynomial time: $O(|U||V||O||B|^2|S|)$, as sketched below:

```
BACKUP$(V_n, B)$
1. $\Gamma^{u,*} \leftarrow \alpha^{u,*}(s) = \sum_v R(s,u,v)\ Pr(v|s)$
2. $\Gamma^{u,v,o} \leftarrow \forall \alpha'_i \in V_n\ \alpha_i^{u,v,o}(s) = \phi\ Pr(v|s) \sum_{s'} Z(s',u,o)\ T(s,u,v,s')\ \alpha'_i(s')$
3. $\Gamma^u_b \leftarrow \Gamma^{u,*} + \sum_{v,o} \arg\max_{\alpha \in \Gamma^{u,v,o}} (\alpha \cdot b)$
4. Return $V_{n+1} \leftarrow \forall b \in B\ \ \arg\max_{\Gamma^u_b, \forall u \in U} (\Gamma^u_b \cdot b)$
```

Time Complexity. Given the set B of sampled beliefs, the cost of solving I-POMDP Lite is divided into two parts: (a) The cost of predicting the mixed strategy of the other agent using nested MDP (2) is $\mathcal{O}\big(kh|S|^2\big)$ (Section 2); (b) To determine the cost of approximately solving I-POMDP Lite with respect to this predicted mixed strategy, since each backup step incurs $\mathcal{O}\big(|U||V||O||B|^2|S|\big)$ time, solving I-POMDP Lite for h steps incurs $\mathcal{O}\big(h|U||V||O||B|^2|S|\big)$ time. By considering $|U|$, $|V|$, and $|O|$ as constants, the cost of solving I-POMDP Lite can be simplified to $\mathcal{O}\big(h|S||B|^2\big)$. Thus, the time complexity of solving I-POMDP Lite is $\mathcal{O}\big(h|S|(k|S| + |B|^2)\big)$, which is much less computationally demanding than the exponential cost of I-PF and I-PBVI (Section 1).

4 Theoretical Analysis

In this section, we prove that I-POMDP Lite inherits convergence, piecewise-linear, and convex properties of POMDP that make it amenable to be solved by existing sampling-based algorithms. More importantly, we show that the performance loss incurred by I-POMDP Lite is linearly bounded by the error of prediction of the other agent's strategy. This result also holds for that of nested MDP policy because I-POMDP Lite reduces to nested MDP under full observability.

Theorem 2 (Convergence)**.** *Let V_∞ be the value function of I-POMDP Lite for infinite time horizon. Then, it is contracting/converging: $\|V_\infty - V_{n+1}\|_\infty \leq \phi \|V_\infty - V_n\|_\infty$.*

Theorem 3 (Piecewise Linearity and Convexity)**.** *The optimal value function V_n can be represented as a finite set of α vectors: $V_n(b) = \max_{\alpha \in V_n} (\alpha \cdot b)$.*

We can prove by induction that the number of α vectors grows exponentially with the length of planning horizon; this explains why deriving the exact I-POMDP Lite policy is intractable in practice.

Definition 1. *Let $\pi^*_{\text{-}t}$ be the true strategy of the other agent -t such that $\pi^*_{\text{-}t}(s,v)$ denotes the true probability $Pr^*(v|s)$ of selecting action $v \in V$ in state $s \in S$ for agent -t. Then, the prediction error is $\epsilon_p \triangleq \max_{v,s} |Pr^*(v|s) - Pr(v|s)|$.*

Definition 2. *Let $R_{\max} \triangleq \max_{s,u,v} R(s,u,v)$ be the maximum value of our agent t's payoffs.*

Theorem 4 (Policy Loss)**.** *The performance loss δ_n incurred by executing I-POMDP Lite policy, induced w.r.t the predicted strategy $\widehat{\pi}^k_{\text{-}t}$ of the other agent -t using nested MDP (as compared to its true strategy $\pi^*_{\text{-}t}$), after n backup steps is linearly bounded by the prediction error ϵ_p :*

$$\delta_n \leq 2\epsilon_p |V| R_{\max} \left[\phi^{n-1} + \frac{1}{1-\phi} \left(1 + \frac{3\phi|O|}{1-\phi} \right) \right] .$$

The above result implies that, by increasing the accuracy of the prediction of the other agent's strategy, the performance of the I-POMDP Lite policy can be proportionally improved. This gives a very strong motivation to seek better and more reliable techniques, other than our proposed nested MDP framework, for intention prediction. The formal proofs of the above theorems are provided in Appendix D.

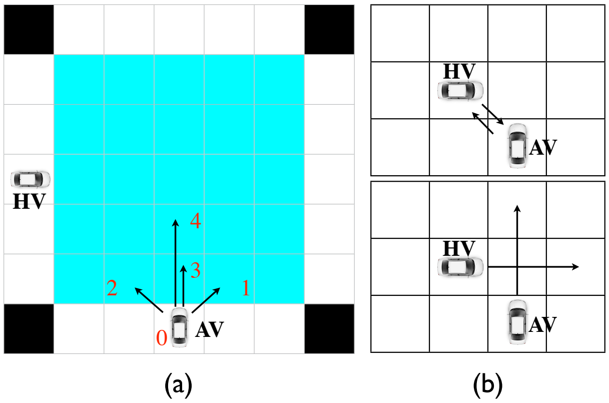

Figure 1: Intersection Navigation: (a) the road intersection modeled as a 7 × 7 grid (the black areas are not passable); and (b) accidents caused by cars crossing trajectories.

5 Experiments and Discussion

This section first evaluates the empirical performance of nested MDP in a practical multi-agent task called *Intersection Navigation for Autonomous Vehicles* (INAV) (Section 5.1), which involves a traffic scenario with multiple cars coming from different directions (North, East, South, West) into an intersection and safely crossing it with minimum delay. Our goal is to implement an intelligent autonomous vehicle (**AV**) that cooperates well with human-driven vehicles (**HV**) to quickly and safely clear an intersection, in the absence of communication. Then, the performance of I-POMDP Lite is evaluated empirically in a series of partially observable stochastic games (POSGs) (Section 5.2). All experiments are run on a Linux server with two 2.2GHz Quad-Core Xeon E5520 processors and 24GB RAM.

5.1 Nested MDP Evaluations

In this task, the road intersection is modeled as a 7×7 grid, as shown in Fig. 1a. The autonomous car (**AV**) starts at the bottom row of the grid and travels North while a human-driven car (**HV**) starts at the leftmost column and travels to the East. Each car has five actions: slow down (0), forward right (1), forward left (2), forward (3) and fast forward (4). Furthermore, it is assumed that 'slow down' has speed level 0, 'forward left', 'forward', and 'forward right' have speed level 1 while 'fast forward' has speed level 2. The difference in speed levels of two consecutive actions should be at most 1. In general, the car is penalized by the delay cost $D > 0$ for each executed action. But, if the joint actions of both cars lead to an accident by crossing trajectories or entering the same cell (Fig. 1b), they are penalized by the accident cost $C > 0$. The goal is to help the autonomous car to safely clear the intersection as fast as possible. So, a smaller value of D/C is desired as it implies a more rational behavior in our agent.

The above scenario is modeled using nested MDP, which requires more than 18000 states. Each state comprises the cells occupied by the cars and their current speed levels. The discount factor ϕ is set to 0.99. The delay and accident costs are hard-coded as $D = 1$ and $C = 100$. Nested MDP is used to predict the mixed strategy of the human driver and our car's optimal policy is computed with respect to this predicted mixed strategy. For evaluation, our car is run through 800 intersection episodes. The human-driven car is scripted with the following rational behavior: the human-driven car probabilistically estimates how likely a particular action will lead to an accident in the next time step, assuming that our car selects actions uniformly. It then forms a distribution over all actions such that most of the probability mass concentrates on actions that least likely lead to an accident. Its next action is selected by sampling from this distribution.

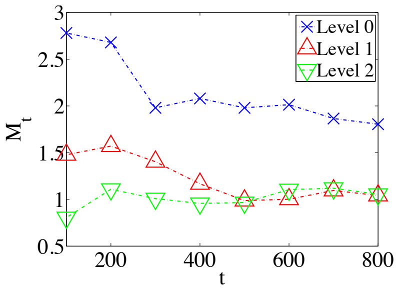

Figure 2: Performance comparison between nested MDPs at reasoning levels 0, 1, and 2.

We compare the performance of nested MDPs at reasoning levels $k = 0, 1, 2$. When $k = 0$, it is equivalent to the traditional MDP policy that treats the other car as environmental noise. During execution, we maintain a running average T_t (over the first t episodes) of the number of actions taken to clear an intersection and the number I_t of intersections experiencing accidents. The average ratio of the empirical delay is defined as $R^d_t = (T_t - T_{\min})/T_{\min} = T_t/T_{\min} - 1$ with $T_{\min} = 3$ (i.e., minimum delay required to clear the intersection). The empirical accident rate is defined as $R^c_t = I_t/t$. The average incurred cost is therefore $M_t = CR^c_t + DR^d_t$. A smaller M_t implies better policy performance.

Fig. 2 shows the results of the performance of the evaluated policies. It can be observed that the M_t curves of nested MDPs at reasoning levels 1 and 2 lie below that of MDP policy (i.e., reasoning level 0). So, nested MDP outperforms MDP. This is expected since our rationality assumption holds: nested MDP's prediction is closer to the human driver's true intention and is thus more informative than the uniformly-distributed human driver's strategy assumed by MDP. Thus, we conclude that nested MDP is effective when the other agent's behavior conforms to our definition of rationality.

5.2 I-POMDP Lite Evaluations

Specifically, we compare the performance of I-POMDP Lite vs. I-POMDP (at reasoning level $k = 1$) players under adversarial environments modeled as zero-sum POSGs. These players are tasked to compete against nested MDP and I-POMDP opponents at reasoning level $k = 0$ (i.e., respectively, MDP and POMDP opponents) whose strategies exactly fit the structural assumptions of I-POMDP Lite (Section 3) and I-POMDP (at $k = 1$), respectively. The I-POMDP player is implemented using I-PBVI which is reported to be the best approximate I-POMDP technique [4]. Each competition consists of 40 stages; the reward/penalty is discounted by 0.95 after each stage. The performance of each player, against its opponent, is measured by averaging its total rewards over 1000 competitions. Our test environment is larger than the benchmark problems in [4]: There are 10 states, 3 actions, and 8 observations for each player. In particular, we let each of the first 6 states be associated with a unique observation with high probability. For the remaining 4 states, every disjoint pair of states is associated with a unique observation with high probability. Hence, the sensing capabilities of I-POMDP Lite and I-POMDP players are significantly weaker than that of the MDP opponent with full observability.

Table 1 shows the results of I-POMDP and I-POMDP Lite players' performance with varying horizon lengths. The observations are as follows: (a) Against a POMDP opponent whose strategy completely favors I-POMDP, both players win by a fair margin and I-POMDP Lite outperforms I-POMDP; (b) against a MDP opponent, I-POMDP suffers a huge loss (i.e., -37.88) as its structural assumption of a POMDP opponent is violated, while I-POMDP Lite wins significantly (i.e., 24.67); and (c) the planning times of I-POMDP Lite and I-POMDP appear to, respectively, grow linearly and exponentially in the horizon length.

Table 1: I-POMDP's and I-POMDP Lite's performance against POMDP and MDP opponents with varying horizon lengths h ($|S| = 10, |A| = 3, |O| = 8$). '$*$' denotes that the program ran out of memory after 10 hours.

	POMDP	MDP	Time (s)	$\|IS'\|$
I-POMDP (h = 2)	13.33±1.75	−37.88±1.74	177.35	66110
I-POMDP (h = 3)	$*$	$*$	$*$	1587010
I-POMDP Lite (h = 1)	15.22±1.81	15.18±1.41	0.02	N.A.
I-POMDP Lite (h = 3)	17.40±1.71	24.23±1.54	0.45	N.A.
I-POMDP Lite (h = 8)	17.42±1.70	24.66±1.54	17.11	N.A.
I-POMDP Lite (h = 10)	17.43±1.70	24.67±1.55	24.38	N.A.

I-POMDP's exponential blow-up in planning time is expected because its bounded interactive state space IS' increases exponentially in the horizon length (i.e., curse of dimensionality), as shown in Table 1. Such a scalability issue is especially critical to large-scale problems. To demonstrate this, Fig. 3b shows the planning time of I-POMDP Lite growing linearly in the horizon length for a large zero-sum POSG with 100 states, 3 actions, and 20 observations for each player; it takes about 6 and

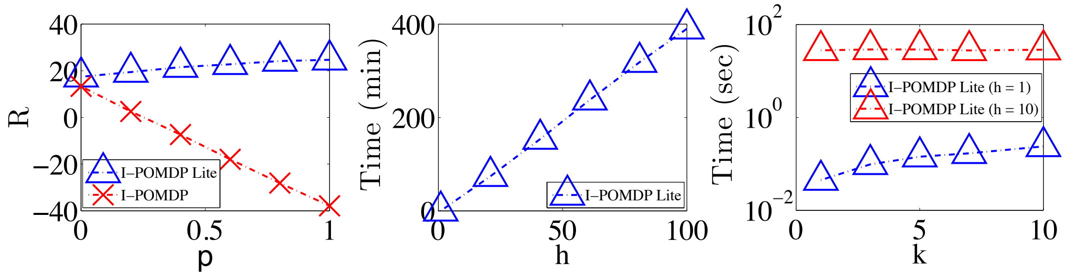

Figure 3: Graphs of (a) performance R of I-POMDP Lite and I-POMDP players against hybrid opponents ($|S| = 10, |A| = 3, |O| = 8$); (b) I-POMDP Lite's planning time vs. horizon length h in a large POSG ($|S| = 100, |A| = 3, |O| = 20$); and (c) I-POMDP Lite's planning time vs. reasoning level k for $h = 1$ and 10 ($|S| = 10, |A| = 3, |O| = 10$).

Table 2: I-POMDP's and I-POMDP Lite's performance against POMDP and MDP opponents with varying horizon lengths h ($|S| = 10, |A| = 3, |O| = 10$). '$*$' denotes that the program ran out of memory after 10 hours.

	POMDP	MDP	Time (s)	$\|IS'\|$
I-POMDP (h = 2)	5.70±1.67	−9.62±1.50	815.28	102410
I-POMDP (h = 3)	$*$	$*$	$*$	3073110
I-POMDP Lite (h = 1)	11.18±1.75	20.25±1.53	0.03	N.A.
I-POMDP Lite (h = 3)	14.89±1.79	27.49±1.53	0.95	N.A.
I-POMDP Lite (h = 8)	14.99±1.79	26.91±1.55	24.10	N.A.
I-POMDP Lite (h = 10)	15.01±1.79	26.91±1.55	33.74	N.A.

$1/2$ hours to plan for 100-step look-ahead. In contrast, I-POMDP fails to even compute its 2-step look-ahead policy within 12 hours.

It may seem surprising that I-POMDP Lite outperforms I-POMDP even when tested against a POMDP opponent whose strategy completely favors I-POMDP. This can be explained by the following reasons: (a) I-POMDP's exponential blow-up in planning time forbids it from planning beyond 3 look-ahead steps, thus degrading its planning performance; (b) as shown in Section 3, the cost of solving I-POMDP Lite is only polynomial in the horizon length and reasoning depth, thus allowing our player to plan with a much longer look-ahead (Fig. 3b) and achieve substantially better planning performance; and (c) with reasonably accurate observation and transition models, Theorem 1 indicates that the strategy of the MDP opponent (i.e., nested MDP at reasoning level 0) is likely to approximate that of the true POMDP opponent closely, thus reducing the degree of violation of I-POMDP Lite's structural assumption of a nested MDP opponent. Such an assumption also seems to make our I-POMDP Lite player overestimate the sensing capability of an unforeseen POMDP opponent and consequently achieve a robust performance against it. On the other hand, the poor performance of I-POMDP against a MDP opponent is expected because I-POMDP's structural assumption of a POMDP opponent is likely to cause its player to underestimate an unforeseen opponent with superior sensing capability (e.g., MDP) and therefore perform badly against it. In contrast, I-POMDP Lite performs significantly better due to its structural assumption of a nested MDP opponent at level 0 which matches the true MDP opponent exactly.

Interestingly, it can be empirically shown that when both players' observations are made more informative than those used in the previous experiment, the performance advantage of I-POMDP Lite over I-POMDP, when tested against a POMDP opponent, increases. To demonstrate this, we modify the previous zero-sum POSG to involve 10 observations (instead of 8) such that every state (instead of a disjoint pair of states) is associated with a unique observation with high probability (i.e., ≥ 0.8); the rest of the probability mass is then uniformly distributed among the other observations. Hence, the sensing capabilities of I-POMDP Lite and I-POMDP players in this experiment are much better than those used in the previous experiment and hence closer to that of the MDP opponent with full observability. Table 2 summarizes the results of I-POMDP Lite's and I-POMDP's performance when tested against the POMDP and MDP opponents in the environment described above.

To further understand how the I-POMDP Lite and I-POMDP players perform when the sensing capability of an unforeseen opponent varies, we set up another adversarial scenario in which both players pit against a hybrid opponent: At each stage, with probability p, the opponent knows the exact state of the game (i.e., its belief is set to be peaked at this known state) and then follows the MDP policy; otherwise, it follows the POMDP policy. So, a higher value of p implies better sensing capability of the opponent. The environment settings are the same as those used in the first experiment, that is, 10 states, 8 observations and 3 actions for each player (Table 1). Fig. 3a shows the results of how the performance, denoted R, of I-POMDP Lite and I-POMDP players vary with p: I-POMDP's performance decreases rapidly as p increases (i.e., opponent's strategy violates I-POMDP's structural assumption more), thus increasing the performance advantage of I-POMDP Lite over I-POMDP. This demonstrates I-POMDP Lite's robust performance when tested against unforeseen opponents whose sensing capabilities violate its structural assumption.

To summarize the above observations, (a) in different partially observable environments where the agents have reasonably accurate observation and transition models, I-POMDP Lite significantly outperforms I-POMDP (Tables 1, 2 and Fig. 3a); and (b) interestingly, it can be observed from Fig. 3a that when the sensing capability of the unforeseen opponent improves, the performance advantage of I-POMDP Lite over I-POMDP increases. These results consistently demonstrate I-POMDP Lite's robust performance against unforeseen opponents with varying sensing capabilities. In contrast, I-POMDP only performs well against opponents whose strategies completely favor it, but its performance is not as good as that of I-POMDP Lite due to its limited horizon length caused by the extensive computational cost of modeling the opponent. Unlike I-POMDP's exponential blow-up in horizon length h and reasoning depth k (Section 1), I-POMDP Lite's processing cost grows linearly in both h (Fig. 3b) and k (Fig. 3c). When $h = 10$, it can be observed from Fig. 3c that I-POMDP Lite's overall processing cost does not change significantly with increasing k because the cost $\mathcal{O}\big(kh|S|^2\big)$ of predicting the other agent's strategy with respect to k is dominated by the cost $\mathcal{O}\big(h|S||B|^2\big)$ of solving I-POMDP Lite for large h (Section 3).

6 Conclusion

This paper proposes the novel nested MDP and I-POMDP Lite frameworks which incorporate the cognitive hierarchy model of games [1] for intention prediction into the normative decision-theoretic POMDP paradigm to address some practical limitations of existing planning frameworks for self-interested MAS such as computational impracticality [4] and restrictive equilibrium theory of agents' behavior [9]. We have theoretically guaranteed that the performance losses incurred by our I-POMDP Lite policies are linearly bounded by the error of intention prediction. We have empirically demonstrated that I-POMDP Lite performs significantly better than the state-of-the-art planning algorithms in partially observable stochastic games. Unlike I-POMDP, I-POMDP Lite's performance is very robust against unforeseen opponents whose sensing capabilities violate the structural assumption (i.e., of a nested MDP opponent) that it has exploited to achieve significant computational gain. In terms of computational efficiency and robustness in planning performance, I-POMDP Lite is thus more practical for use in larger-scale problems.

References

[1] C. F. Camerer, T. H. Ho, and J. K. Chong. A cognitive hierarchy model of games. *Quarterly J. Economics*, 119(3):861–898, 2004.

[2] G. Chalkiadakis and C. Boutilier. Coordination in multiagent reinforcement learning: A Bayesian approach. In *Proc. AAMAS*, pages 709–716, 2003.

[3] P. Doshi and P. Gmytrasiewicz. Monte Carlo sampling methods for approximating interactive POMDPs. *JAIR*, pages 297–337, 2009.

[4] P. Doshi and D. Perez. Generalized point based value iteration for interactive POMDPs. In *Proc. AAAI*, pages 63–68, 2008.

[5] G. Gigerenzer and R. Selten. *Bounded Rationality*. MIT Press, 2002.

[6] P. J. Gmytrasiewicz and P. Doshi. A framework for sequential planning in multi-agent settings. *JAIR*, 24:49–79, 2005.

[7] E. A. Hansen, D. S. Bernstein, and S. Zilberstein. Dynamic programming for partially observable stochastic games. In *Proc. AAAI*, pages 709–715, 2004.

[8] T. N. Hoang and K. H. Low. A general framework for interacting Bayes-optimally with self-interested agents using arbitrary parametric model and model prior. In *Proc. IJCAI*, 2013.

[9] J. Hu and M. P. Wellman. Multi-agent reinforcement learning: Theoretical framework and an algorithm. In *Proc. ICML*, pages 242–250, 1998.

[10] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proc. ICML*, pages 157–163, 1994.

[11] R. Nair and M. Tambe. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings. In *Proc. IJCAI*, pages 705–711, 2003.

[12] B. Ng, C. Meyers, K. Boakye, and J. Nitao. Towards applying interactive POMDPs to real-world adversary modeling. In *Proc. IAAI*, pages 1814–1820, 2010.

[13] J. Pineau, G. Gordon, and S. Thrun. Point-based value iteration: An anytime algorithm for POMDPs. In *Proc. IJCAI*, pages 1025–1032, 2003.

[14] B. Rathnasabapathy, P. Doshi, and P. Gmytrasiewicz. Exact solutions of interactive POMDPs using behavioral equivalence. In *Proc. AAMAS*, pages 1025–1032, 2006.

[15] S. Seuken and S. Zilberstein. Memory-bounded dynamic programming for DEC-POMDPs. In *Proc. IJCAI*, pages 2009–2015, 2007.

[16] M. T. J. Spaan, F. A. Oliehoek, and C. Amato. Scaling up optimal heuristic search in DEC-POMDPs via incremental expansion. In *Proc. IJCAI*, pages 2027–2032, 2011.

A Proof Sketch of Theorem 1

Theorem 1. Let $\widehat{Q}^n_{\text{-}t}(s,v) \triangleq |U|^{-1}\sum_{u\in U} Q^{0,n}_{\text{-}t}(s,v,u)$ and $\widehat{Q}^n_{\text{-}t}(b,v)$ denote n-step-to-go values of selecting action $v \in V$ in state $s \in S$ and belief b, respectively, for the other agent -t using nested MDP and I-POMDP at reasoning level 0 (i.e., MDP and POMDP). If $b(s) \geq 1-\epsilon$ and

$$\forall\,(s,v,u)\,\exists\,(s',o)\;Pr(s'|s,v,u) \geq 1-\frac{\epsilon}{2}\;\wedge\;Pr(o|s,v) \geq 1-\frac{\epsilon}{2}$$

for some $\epsilon \geq 0$, then

$$\left|\widehat{Q}^n_{\text{-}t}(s,v) - \widehat{Q}^n_{\text{-}t}(b,v)\right| \leq \epsilon\,\mathcal{O}\!\left(\frac{R_{\max}-R_{\min}}{1-\phi}\right) \tag{8}$$

where $R_{\max}$ and $R_{\min}$ denote agent -t's maximum and minimum immediate payoffs, respectively.

Proof: Assuming that (8) holds with n (it is trivial to verify that it holds with $n=0$), we need to prove that it also holds with $n+1$. Define $L^{n+1}_{\text{-}t}(s,v)$ as the optimal expected utility of the other agent if it knows that the first state is s and executes v in the first step. From the second step, it only observes the environmental state partially (i.e., same as the POMDP agent). Then,

$$\begin{aligned}\left|\widehat{Q}^{n+1}_{\text{-}t}(s,v) - \widehat{Q}^{n+1}_{\text{-}t}(b,v)\right| &\leq \left|\widehat{Q}^{n+1}_{\text{-}t}(s,v) - L^{n+1}_{\text{-}t}(s,v)\right| \\ &+ \left|L^{n+1}_{\text{-}t}(s,v) - \widehat{Q}^{n+1}_{\text{-}t}(b,v)\right|\end{aligned}$$

I. Note that $\widehat{Q}^{n+1}_{\text{-}t}(b,v) = \sum_{s'}\alpha^*_v(s')b(s')$ due to piecewise linearity and convexity and $L^{n+1}_{\text{-}t}(s,v) = \alpha^*_v(s)$, by definition. Assuming that $L^{n+1}_{\text{-}t}(s,v) \geq \widehat{Q}^{n+1}_{\text{-}t}(b,v)$, the second term in the right-hand side of the above inequality is linearly bounded w.r.t ϵ, as shown below:

$$\begin{aligned}L^{n+1}_{\text{-}t}(s,v) - \widehat{Q}^{n+1}_{\text{-}t}(b,v) &\leq (\alpha^*_v(s) - \min_{s'\neq s}\alpha^*_v(s'))(1-b(s)) \\ &\leq \epsilon(\alpha^*_v(s) - \min_{s'\neq s}\alpha^*_v(s'))\end{aligned}$$

Similarly, if $L^{n+1}_{\text{-}t}(s,v) \leq \widehat{Q}^{n+1}_{\text{-}t}(b,v)$, we can instead show that $\widehat{Q}^{n+1}_{\text{-}t}(b,v) - L^{n+1}_{\text{-}t}(s,v) \leq \epsilon(\max_{s'\neq s}\alpha^*_v(s') - \alpha^*_v(s))$. Thus, it follows that

$$\begin{aligned}\left|L^{n+1}_{\text{-}t}(s,v) - \widehat{Q}^{n+1}_{\text{-}t}(b,v)\right| &\leq \epsilon\left(\max_{s'}\alpha^*_v(s') - \min_{s'}\alpha^*_v(s')\right) \\ &\leq \epsilon\,\mathcal{O}\!\left(\frac{R_{\max}-R_{\min}}{1-\phi}\right)\end{aligned} \tag{9}$$

II. From our definitions of $\widehat{Q}^{n+1}_{\text{-}t}(s,v)$ and $L^{n+1}_{\text{-}t}(s,v)$, it directly follows that

$$\widehat{Q}^{n+1}_{\text{-}t}(s,v) = \sum_u \frac{1}{|U|}\left(R_{\text{-}t}(s,v,u) + \phi\sum_{s'} T^{u,v}_s(s')\widehat{V}^n_{\text{-}t}(s')\right)$$

where $\widehat{V}^n_{\text{-}t}(s') \triangleq \max_{v'}\widehat{Q}^n_{\text{-}t}(s',v')$, $R_{\text{-}t}(s,v,u)$ is the other agent's immediate payoff, and $T^{u,v}_s(s') \triangleq Pr(s'|s,v,u)$. Similarly, $L^{n+1}_{\text{-}t}(s,v)$ can be expressed as

$$L^{n+1}_{\text{-}t}(s,v) = \sum_u \frac{1}{|U|}\left(R_{\text{-}t}(s,v,u) + \phi\sum_{s',o} T^{u,v}_s(s')Z^v_{s'}(o)\widehat{V}^n_{\text{-}t}(b^{v,o}_u)\right)$$

where $\widehat{V}^n_{\text{-}t}(b^{v,o}_u) \triangleq \max_{v'}\widehat{Q}^n_{\text{-}t}(b^{v,o}_u,v')$, $Z^v_{s'}(o) \triangleq Pr(o|s',v)$, and $b^{v,o}_u(s') = \dfrac{Z^v_{s'}(o)T^{u,v}_s(s')}{\sum_{s''}Z^v_{s''}(o)T^{u,v}_s(s'')}$[4]. Then,

$$\left|\widehat{Q}^{n+1}_{\text{-}t}(s,v) - L^{n+1}_{\text{-}t}(s,v)\right| \leq \phi\sum_u \frac{1}{|U|}F(u) \tag{10}$$

[4] By definition of $L^{n+1}_{\text{-}t}(s,v)$, the previous state s is known exactly. So, we do not need to take the average w.r.t $b(s)$.

where $F(u) \triangleq \sum_{s'} T^{u,v}_s(s') \left|\widehat{V}^n_{\text{-}t}(s') - \sum_o Z^v_{s'}(o)\widehat{V}^n_{\text{-}t}(b^{v,o}_u)\right|$. Let s^* be the state at which $T^{u,v}_s(s^*) \geq 1 - \dfrac{\epsilon}{2}$, we have

$$\begin{aligned}
F(u) \quad &\leq \quad \sum_{s' \neq s^*} T^{u,v}_s(s') \left|\widehat{V}^n_{\text{-}t}(s') - \sum_o Z^v_{s'}(o)\widehat{V}^n_{\text{-}t}(b^{v,o}_u)\right| \\
&+ \quad T^{u,v}_s(s^*) \left|\widehat{V}^n_{\text{-}t}(s^*) - \sum_o Z^v_{s^*}(o)\widehat{V}^n_{\text{-}t}(b^{v,o}_u)\right| \qquad (11) \\
&\leq \quad \mathcal{O}\left(\frac{R_{\max} - R_{\min}}{1-\phi}\right) \sum_{s' \neq s^*} T^{u,v}_s(s') \\
&+ \quad T^{u,v}_s(s^*) \left|\widehat{V}^n_{\text{-}t}(s^*) - \sum_o Z^v_{s^*}(o)\widehat{V}^n_{\text{-}t}(b^{v,o}_u)\right| \qquad (12)
\end{aligned}$$

Since $T^{u,v}_s(s^*) \geq 1 - \dfrac{\epsilon}{2}$ and $\sum_{s' \neq s^*} T^{u,v}_s(s') = 1 - T^{u,v}_s(s^*)$, it follows directly that the first term in the right-hand side of (12) is bounded by $\dfrac{\epsilon}{2}\mathcal{O}((R_{\max} - R_{\min})/(1-\phi))$. Then, let $G(s^*, u) \triangleq T^{u,v}_s(s^*) \left|\widehat{V}^n_{\text{-}t}(s^*) - \sum_o Z^v_{s^*}(o)\widehat{V}^n_{\text{-}t}(b^{v,o}_u)\right|$ and let o^* be the value at which $Z^v_{s^*}(o^*) \geq 1 - \dfrac{\epsilon}{2}$.

$$\begin{aligned}
G(s^*, u) \quad &\leq \quad \left|\widehat{V}^n_{\text{-}t}(s^*) - \widehat{V}^n_{\text{-}t}(b^{v,o^*}_u)\right| \\
&+ \quad \left|\widehat{V}^n_{\text{-}t}(b^{v,o^*}_u) - \sum_o Z^v_{s^*}(o)\widehat{V}^n_{\text{-}t}(b^{v,o}_u)\right| \qquad (13)
\end{aligned}$$

Since the first term on the right-hand side of (13) is trivially bounded by $\max_{v'} \left|\widehat{Q}^n_{\text{-}t}(s^*, v') - \widehat{Q}^n_{\text{-}t}(b^{v,o^*}_u, v')\right|$ and the posterior belief $b^{v,o^*}_u(s^*) \geq T^{u,v}_s(s^*)Z^v_{s^*}(o^*) \geq \left(1 - \dfrac{\epsilon}{2}\right)^2 \geq 1 - \epsilon$, we can bound it by $\epsilon\mathcal{O}((R_{\max} - R_{\min})/(1-\phi))$ using our inductive assumption. Also, using the same argument in part **I**, the second term is bounded by $\dfrac{\epsilon}{2}\mathcal{O}\left(\dfrac{R_{\max} - R_{\min}}{1-\phi}\right)$. This implies $G(s^*, u)$ and hence $F(u)$ are bounded by $\epsilon\mathcal{O}\left(\dfrac{R_{\max} - R_{\min}}{1-\phi}\right)$. Plugging $F(u)$ into (10), we can prove that $\left|\widehat{Q}^{n+1}_{\text{-}t}(s, v) - L^{n+1}_{\text{-}t}(s, v)\right| \leq \epsilon\mathcal{O}\left(\dfrac{R_{\max} - R_{\min}}{1-\phi}\right)$.

III. Putting the results in parts **I** and **II** together, it follows that (8) holds for $n + 1$ as well.

B Empirical Evaluations for Nested MDP

This section evaluates the performance of nested MDP in a two-player, zero-sum Markov game modeled after soccer. In particular, the soccer game introduced in [Littman, 1994] is played on a 4×5 grid, as shown in Fig. 4. At any moment, the two players A and B occupy different squares and have five actions: top, down, left, right, and stand. Only one of them has the ball. Once they have selected their actions, the moves are executed in random order. When the player with the ball moves into its assigned goal (left and right shaded squares for A and B, respectively), it gets one point and its opponent loses a point and vice versa. The game is then reset to the initial configuration shown in Fig. 4 with the ball possession going to either one at random. When a player steps into the square occupied by its opponent, the ball possession goes to its opponent and the move does not take place. The discount factor is set to 0.9, which makes scoring sooner better than scoring later.

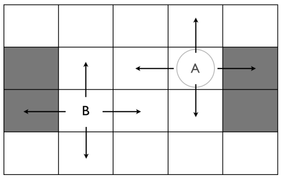

Figure 4: Soccer game.

The performance of nested MDP player at level $k = 1$ is compared against that of the state-of-the-art Nash-Q [Hu and Wellman, 1998] and the MDP planning players. These players are tasked to compete against opponents employing Nash-Q and MDP planning, a random policy selecting actions uniformly, and a hand-built policy specifying simple scoring and blocking rules:

Scoring. This set of tactics applies when the agent possesses the ball. At any step, if the agent is not on the right track (i.e., its current row goes straight to a goal), it will consider switching to the closest track while avoiding its opponent. Otherwise, it will try to reach the goal as fast as possible without losing the ball to the defending opponent.

Specifically, when the agent is on the right track, it will just keep moving towards the goal if the opponent is not blocking. Otherwise, it will only move forward with a certain probability which is set to be proportional to its distance from the opponent. If the agent decides not to move forward, it will consider staying still with a small probability to maintain the ball. If not, it either moves left or right with equal probability to keep away from the opponent.

When the agent is not on the right track, it simply moves to the closest track if the defending opponent is left behind or too far ahead (i.e., more than two steps ahead). Otherwise, if the opponent is close up ahead, it will consider either moving forward (with the hope of getting past its opponent) or switching to the right track with certain probabilities. Again, we make the agent's probability of moving forward proportional to its distance from the opponent.

Blocking. When the agent loses the ball, the following tactics apply: The agent will chase after the opponent (who is rushing towards its corresponding goal) if it is left behind. Otherwise, it either switches to the same track (i.e., row) as the opponent or stays still if their tracks are currently the same. That is, the agent always tries to stand in the way of the opponent, which appears to be a very effective defending strategy [Littman, 1994].

Each competition between a player and its opponent consists of a number of games, each of which ends immediately after either the first goal or being declared as a draw with 0.1 probability at every step. But, each competition only ends after 10000 games that do not end as a draw.

Table 3 shows the results of the performance of the tested players. The observations are as follows:

Players Opponents	**Nash-Q**	**Nested MDP**	**MDP**
Nash-Q	4970 (173051)	4955 (140867)	2393 (74545)
MDP	7602 (74458)	7558 (53946)	4983 (37277)
Random	9771 (128388)	9936 (96865)	9770 (77992)
Hand-built	5575 (130843)	6069 (99698)	2880 (56854)

Table 3: Comparison of number of goals scored in a competition by different players: the number of moves taken to complete 10000 goals are specified in brackets.

(a) Against a Nash-Q opponent, the nested MDP player's performance is comparable to that of the Nash-Q player. In theory, the best strategy against a Nash-Q opponent should be a corresponding one in the same Nash equilibrium. This implies the nested MDP player achieves a near-optimal policy against the Nash-Q opponent. Intuitively, although the nested MDP player does not expect the opponent to play Nash strategy, it can roughly predict which of its opponent's moves would potentially lead to its loss and how likely these moves will be executed. Exploiting this knowledge, it plans a policy that rationally trades off between defense and offense, thereby achieving near-optimal performance against the Nash-Q opponent.

(b) Against a MDP opponent, nested MDP and Nash-Q players' performance are superior to that of the MDP player who wrongly assumes the random action selection by the opponent, thereby leading to an offensive policy with weak defense. Notably, while nested MDP and Nash-Q players' performance are comparable in this case, the nested MDP player completed the competition significantly faster (i.e., fewer moves taken) than the Nash-Q player.

(c) Against a random-policy opponent, Nash-Q and MDP players' performance are almost similar while the nested MDP player performs significantly better. Upon close examination, we observe that, in such a two-player zero-sum game, the Nash-Q strategy coincides with the maximin strategy. So, its corresponding policy is too risk-averse and fails to fully exploit the weaknesses of the random-policy opponent, thereby losing some chances of winning. In contrast, the MDP player performs well because its offensive strategy is generally effective against the weak random-policy opponent. Lastly, since the nested MDP player always has a better balance between defense and offense, its performance is the best among the players. These behaviors can be observed from the number of moves the players took to finish the competition against the same opponent.

(d) Against a hand-built policy opponent, Nash-Q and nested MDP players win more than half the time while the MDP player performs badly due to its ignorance of the opponent's intention, as shown in previous experiments. In particular, the nested MDP player significantly outperforms the Nash-Q player: the nested MDP player, while planning, predicts the opponent's intention by recursively reasoning in its place, thereby recognizing its opponent's critical moves that can potentially lead to its loss. Exploiting this knowledge, the nested MDP player usually has a better trade-off between offense and defense than the Nash-Q player who only focuses on the worst-case situations, thereby losing chances to score if its opponent does not make moves that lead to such situations.

Essentially, these observations show that nested MDP (1) significantly outperforms MDP player (against all opponents) and Nash-Q player (against Random and Hand-built opponents) and (2) performs comparably to Nash-Q player (against Nash-Q and MDP opponents). This suggests that nested MDP is a better planning framework under fully observable settings as compared against the previous ones such as Nash-Q and MDP planning.

C Absolute Continuity Condition of Interactive Beliefs for I-PBVI

In this section, we provide a more detailed discussion on the *Absolute Continuity Condition* (ACC) of interactive beliefs [Doshi and Perez, 2008], which is crucial to the mathematical soundness and hence the feasibility of using I-PBVI. Indeed, we have ensured that this condition is met in all of our experiments so that I-PBVI is not put at a disadvantage against I-POMDP Lite. Intuitively, I-PBVI aims to make the planning of I-POMDP tractable by constraining the infinite interactive state space $IS = S \times \Delta(S)$ to a finite space $IS' = S \times \text{Reach}(B, h)$ bounded with respect to the length h of planning horizon. Since $\text{Reach}(B, h)$ includes all beliefs that can be reached from B by following a certain action-observation history of length h, if the true initial belief of the other agent is included in B, then all its reachable beliefs within h steps of interaction are included in $\text{Reach}(B, h)$ as well. This essentially means that IS' includes all interactive states of IS that can be assigned a non-zero probability within h steps of interaction and hence guarantees the correctness of the interactive belief update on IS'.

On the other hand, if B does not include the other agent's true initial belief, it is possible that the other agent actually reaches an interactive state not included in $\text{Reach}(B, h)$ within h steps of interaction. Consequently, this causes the loss of probability mass while updating the interactive belief: if the action-observation history assigns a non-zero probability mass to an interactive state not included in IS', then this mass will be lost because the interactive belief update step only considers those included in IS'. As the interaction proceeds, the probability mass will disappear gradually until the whole probability mass, which should sum to 1, is completely lost (i.e., the actual action-observation history assigns zero probability to all interactive states included in IS'). At this point, the agent will become indifferent between the actions and thus perform terribly. We have also verified this by trying to exclude the other agent's true initial belief from B: after a few steps, our agent's interactive belief disappears and its performance starts to degrade (i.e., getting negative rewards).

So, the ACC condition simply requires that the other agent's true initial belief should be included in B. However, the specification of B is often complicated: if B is made too large just to increase the chance of including the other agent's true initial belief, the planning process may become computationally intractable since $|IS'| = |S| \times |\text{Reach}(B, h)| = \mathcal{O}\big(|S||B||U|^h|V|^h|O|^h\big)$. Otherwise, there is a need for highly accurate and informative prior knowledge that can satisfy the ACC condition with a smaller-sized B. While such detailed prior knowledge of the other agent is assumed in all our experiments as well as the benchmark problems reported in [Doshi and Perez, 2008], one should be cautioned that it cannot be easily accessed nor processed in real-world situations (e.g., involving humans). This is another obstacle that limits I-PBVI's practical use.

D Proofs

We hereby present our formal proofs for the theorems stated in Section 4. To increase the readability of this section, we structure our proof in three separate sections:

- **Definitions** (Section D.1) – we introduce the conventional definitions used in this analysis, some of which are previously stated in Section 4.
- **Intermediate Results** (Section D.2) – we state and prove several intermediate results, which are necessary to prove our main theorems.
- **Main Theorems** (Section D.3) – finally, we formally derive the main results stated in Section 4.

D.1 Definitions

Definition 1. Let $R_{\max} \triangleq \max_{s,u,v} R(s,u,v)$ be the maximum value of our agent's payoffs and let $\pi^*_{\text{-}t}$ be the true mixed strategy of the other agent -t such that $\pi^*_{\text{-}t}(s,v)$ denotes the true probability $Pr^*(v|s)$ of agent -t selecting action $v \in V$ in state $s \in S$. Then, the prediction error is

$$\epsilon_p \triangleq \max_{v,s} |Pr^*(v|s) - Pr(v|s)| \ .$$

Definition 2. Given the true mixed strategy $\pi^*_{\text{-}t}(s,v) \triangleq Pr^*(v|s)$ of the agent -t, let us denote $Pr^*(v,o|b,u)$, $B^*(b,u,v,o)$ and $V^*_n(b)$ as the belief-state observation model, belief-state update function and optimal value function over time horizon n computed with respect to $Pr^*(v|s)$.

Definition 3. Let $V_n(b)$ denotes the optimal value function computed with respect to the prediction $Pr(v|s)$. The difference between $V_n(b)$ and the true optimal value function $V^*_n(b)$ is $\delta_n \triangleq \max_b |V^*_n(b) - V_n(b)|$.

Definition 4. Let $Q^*_n(b,u)$ and $Q_n(b,u)$ denote the corresponding Q-functions of $V^*_n(b)$ and $V_n(b)$:

$$\begin{aligned} V^*(b) &= \max_u Q^*_n(b,u) \\ V_n(b) &= \max_u Q_n(b,u) \ . \end{aligned}$$

Consequently, we define the optimal policies π^* and π induced from V^*_n and V_n, respectively, as:

$$\begin{aligned} \pi^*(b) &= \arg\max_u Q^*_n(b,u) \\ \pi(b) &= \arg\max_u Q_n(b,u) \ . \end{aligned}$$

Definition 5. Let $J^*_n(b)$ and $J_n(b)$ denote the expected total rewards if, starting from belief b, we follow the optimal policy π^* and the induced policy π for n steps, respectively. Thus, we have:

$$\begin{aligned} J^*_n(b) &= \sum_{s,v} b(s) Pr^*(v|s) R(s,\pi^*(b),v) \\ &+ \phi \sum_{v,o} Pr^*(v,o|b,\pi^*(b)) J^*_{n-1}(B^*(b,\pi^*(b),v,o)) \\ J_n(b) &= \sum_{s,v} b(s) Pr^*(v|s) R(s,\pi(b),v) \\ &+ \phi \sum_{v,o} Pr^*(v,o|b,\pi(b)) J_{n-1}(B^*(b,\pi(b),v,o)) \ . \end{aligned}$$

Consequently, it can be verified by the definition of the value function V^*_n that $J^*_n \equiv V^*_n$.

D.2 Immediate Results

Lemma 1. Given a prediction $\pi_{\text{-}t}(s,v) \triangleq Pr(v|s)$ of the other agent -t's mixed strategy, the difference between the optimal value function V_n and the true optimal value function V^*_n of our agent t, with respect to agent -t's true mixed strategy $\pi^*_{\text{-}t}(s,v) \triangleq Pr^*(v|s)$, is bounded by:

$$\delta_n \leq \max_b \max_u |Q^*_n(b,u) - Q_n(b,u)| \ . \tag{14}$$

Proof: Let us denote $u' = \arg\max_u Q_n(b,u)$ and $u^* = \arg\max_u Q^*_n(b,u)$. Assume that $V^*_n(b) \leq V_n(b)$, we have

$$\begin{aligned} |V^*_n(b) - V_n(b)| &= V_n(b) - V^*_n(b) = Q_n(b,u') - \max_u Q^*_n(b,u) \\ &\leq Q_n(b,u') - Q^*(b,u') = |Q_n(b,u') - Q^*_n(b,u')| \end{aligned} \tag{15}$$

Similarly, if $V_n(b) \leq V_n^*(b)$ we also have

$$|V_n^*(b) - V_n(b)| \quad \leq \quad |Q_n^*(b,u^*) - Q_n(b,u^*)| \tag{16}$$

Therefore, from inequalities (15) and (16), we have

$$\begin{aligned} |V_n^*(b) - V_n(b)| &\leq \max\left(|Q_n(b,u') - Q_n^*(b,u')|, |Q_n^*(b,u^*) - Q_n(b,u^*)|\right) \\ &\leq \max_u |Q_n^*(b,u) - Q_n(b,u)| \end{aligned} \tag{17}$$

From (17) and Definition 3, we have $\delta_n \leq \max_b \max_u |Q_n^*(b,u) - Q_n(b,u)|$ □

Lemma 2. Suppose that at current belief b, our agent (i.e., agent t) executes action u. We defined the expected immediate payoffs with respect to our predictive distribution $\pi_{\text{-}t}(s,v) \triangleq Pr(v|s)$ and the true mixed strategy $\pi^*_{\text{-}t}(s,v) \triangleq Pr^*(v|s)$ of agent -t as

$$\begin{aligned} R^*(b,u) &= \sum_s b(s) \sum_v R(s,u,v) Pr^*(v|s) \\ R(b,u) &= \sum_s b(s) \sum_v R(s,u,v) Pr(v|s)\ . \end{aligned}$$

The difference between $R^*(b,u)$ and $R(b,u)$ is linearly bounded by the prediction error ϵ_p:

$$|R^*(b,u) - R(b,u)| \leq \epsilon_p |V| R_{\max}\ .$$

Proof: We have

$$\begin{aligned} |R^*(b,u) - R(b,u)| &\leq \sum_s b(s) \sum_v R(s,u,v) |Pr^*(v|s) - Pr(v|s)| \\ &\leq \sum_s b(s) \sum_v R(s,u,v)\epsilon_p \leq \sum_s b(s)|V| R_{\max}\epsilon_p \\ &\leq |V| R_{\max}\epsilon_p\ . \end{aligned} \tag{18}$$

□

Lemma 3. The difference between the belief-state observation models $Pr^*(v,o|b,u)$ and $Pr(v,o|b,u)$ with respect to agent -t's true mixed strategy $\pi^*_{\text{-}t}(s,v) \triangleq Pr^*(v|s)$ and our predictive distribution $\pi_{\text{-}t}(s,v) \triangleq Pr(v|s)$ is bounded by the prediction error ϵ_p:

$$|Pr^*(v,o|b,u) - Pr(v,o|b,u)| \leq \epsilon_p\ .$$

Proof: We have

$$\begin{aligned} |Pr^*(v,o|b,u) - Pr(v,o|b,u)| &\leq \sum_{s'} Z(s',u,o) \sum_s T(s,u,v,s') b(s) |Pr^*(v|s) - Pr(v|s)| \\ &\leq \epsilon_p \sum_s b(s) \sum_{s'} Z(s',u,o) T(s,u,v,s') \\ &\leq \epsilon_p \sum_s b(s) \sum_{s'} T(s,u,v,s') = \epsilon_p \sum_s b(s) = \epsilon_p\ . \end{aligned} \tag{19}$$

□

Lemma 4. For any two beliefs b and b', if $\|b - b'\|_1 \leq \delta$, we have $|V(b) - V(b')| \leq \frac{R_{\max}}{1-\phi}\delta$. (Lipschitz condition)

Proof: For any two beliefs b and b', let $\alpha = \arg\max_{\alpha \in V}(\alpha b)$ and $\alpha' = \arg\max_{\alpha \in V}(\alpha b')$. Without loss of generality, assuming that $V(b') \leq V(b)$, we have

$$\begin{aligned}
|V(b) - V(b')| &= \alpha b - \alpha' b' \\
&\leq \alpha b - \alpha b' = \alpha(b - b') \\
&= \sum_s \alpha(s)(b(s) - b'(s)) \leq \sum_s \alpha(s)|(b(s) - b'(s)| \\
&\leq \max_s \alpha(s) \sum_s |b(s) - b'(s)| \leq \max_s \alpha(s)\delta \\
&\leq \frac{R_{\max}}{1-\phi}\delta \; . \qquad (20)
\end{aligned}$$

The last step follows from the fact that each component $\alpha(s)$ of an α vector is basically an expected total reward (if the initial state is s and the agent follows the optimal policy) which is always bounded by $\frac{R_{max}}{1-\phi}$ □

Lemma 5. Let us define the unnormalized belief update function $F(b, u, v, o)$ of $B(b, u, v, o)$ as the followings:

$$\begin{aligned}
F(b, u, v, o)(s') &= Z(s', u, o) \sum_s T(s, u, v, s') Pr(v|s) b(s) \\
&= B(b, u, v, o) Pr(v, o|b, u) \; .
\end{aligned}$$

The norm-1 distance between the unnormalized belief update functions $F^*(b, u, v, o)$ and $F(b, u, v, o)$ of $B^*(b, u, v, o)$ and $B(b, u, v, o)$, respectively, is at most ϵ_p.

Proof: Let us shortly write $F(b, u, v, o)$ and $F^*(b, u, v, o)$ as f' and f^*, respectively. We have

$$\begin{aligned}
\|f' - f^*\|_1 &= \sum_{s'} Z(s', u, o) \sum_s T(s, u, v, s') b(s) |Pr(v|s) - Pr^*(v|s)| \\
&\leq \epsilon_p \sum_{s'} Z(s', u, o) \sum_s T(s, u, v, s') b(s) \\
&= \epsilon_p \sum_s b(s) \sum_{s'} Z(s', u, o) T(s, u, v, s') \\
&\leq \epsilon_p \sum_s b(s) \sum_{s'} T(s, u, v, s') \\
&= \epsilon_p \sum_s b(s) = \epsilon_p \; . \qquad (21)
\end{aligned}$$

□

Proposition 1. For all values $b \in B, u \in U, v \in V, o \in O$, we have

$$Pr^*(v, o|b, u) \|B^*(b, u, v, o) - B(b, u, v, o)\|_1 \leq 2\epsilon_p \; .$$

Proof: Let b^* and b' denote $B^*(b, u, v, o)$ and $B(b, u, v, o)$, respectively. Also, let f^* and f' be unnormalized version of b^* and b'. Finally, let A^* and A denote the belief-state observation probability

$Pr^*(v, o|b, u)$ and $Pr(v, o|b, u)$. Thus, we have

$$\begin{aligned}
A^*\|b^* - b'\|_1 &= A^* \sum_{s'} \left| \frac{f^*(s')}{A^*} - \frac{f'(s')}{A} \right| \\
&= \frac{1}{A} \sum_{s'} |f^*(s')A - f'(s')A^*| \\
&= \frac{1}{A} \sum_{s'} |A(f^*(s') - f'(s')) + f'(s')(A - A^*)| \\
&\leq \frac{1}{A} \sum_{s'} (A|f^*(s') - f'(s')| + f'(s')|A - A^*|) \\
&\leq \frac{1}{A} \sum_{s'} (A|f^*(s') - f'(s')| + f'(s')\epsilon_p) \text{ (Lemma 3)} \\
&= \sum_{s'} |f^*(s') - f'(s')| + \epsilon_p \frac{1}{A} \sum_{s'} f'(s') \\
&= \sum_{s'} |f^*(s') - f'(s')| + \epsilon_p \frac{1}{A} \sum_{s'} Ab'(s') \text{ (Def. of } f'(.)) \\
&= \sum_{s'} |f^*(s') - f'(s')| + \epsilon_p \\
&\leq \epsilon_p + \epsilon_p = 2\epsilon_p \text{ (Lemma 5)} \qquad (22)
\end{aligned}$$

□

Proposition 2. Given a predictive distribution $\pi_{\text{-}t}(s, v) \triangleq Pr(v|s)$ of agent -t's mixed strategy, we have the following inequality:

$$|V^*(B^*(b, u, v, o))Pr^*(v, o|b, u) - V^*(B(b, u, v, o))Pr(v, o|b, u)| \leq 3\epsilon_p \frac{R_{\max}}{1 - \phi}$$

Proof: Let b^* and b' denote $B^*(b, u, v, o)$ and $B(b, u, v, o)$, respectively. Also, let A^* and A denote the probability $Pr^*(v, o|b, u)$ and $Pr(v, o|b, u)$. Thus, we have:

$$\begin{aligned}
|V^*(b^*)A^* - V^*(b')A| &\leq |V^*(b^*)A^* - V^*(b')A^*| + |V^*(b')A^* - V^*(b')A| \\
&= A^*\,|V^*(b^*) - V^*(b')| + V^*(b')\,|A^* - A| \\
&\leq A^* \frac{R_{\max}}{1 - \phi} \|b^* - b'\|_1 + V^*(b')\,|A^* - A| \text{ (Lemma 4)} \\
&= \frac{R_{\max}}{1 - \phi} A^* \|b^* - b'\|_1 + V^*(b')\,|A^* - A| \\
&\leq 2\epsilon_p \frac{R_{\max}}{1 - \phi} + V^*(b')\,|A^* - A| \text{ (Proposition 1)} \\
&\leq 2\epsilon_p \frac{R_{\max}}{1 - \phi} + V^*(b')\epsilon_p \text{ (Lemma 3)} \\
&\leq 2\epsilon_p \frac{R_{\max}}{1 - \phi} + \epsilon_p \frac{R_{\max}}{1 - \phi} = 3\epsilon_p \frac{R_{\max}}{1 - \phi}\,. \qquad (23)
\end{aligned}$$

The last step follows from the fact that the value function $V^*(b')$ is essentially the expected total reward if the agent follows the optimal policy from the initial belief. Thus, it is trivially bounded by $\frac{R_{\max}}{1-\phi}$ □

Proposition 3. For all values $b \in B, u \in U$, we have:

$$|Q^*_n(b, u) - Q_n(b, u)| \leq \phi\delta_{n-1} + \epsilon_p |V| R_{\max} \left(1 + 3\phi \frac{|O|}{1 - \phi}\right) .$$

Proof: Let b^* and b' denote $B^*(b, u, v, o)$ and $B(b, u, v, o)$, respectively. Further, let us define

$$L_n(b, u) \quad = \quad R(b, u) + \phi \sum_v \sum_o V^*_{n-1}(b') Pr(v, o|b, u) \; .$$

First, we prove that $|L_n(b, u) - Q_n(b, u)| \leq \phi\delta_{n-1}$:

$$\begin{aligned} |L_n(b, u) - Q_n(b, u)| &\leq \phi \sum_{v,o} Pr(v, o|b, u) |V^*_{n-1}(b') - V_{n-1}(b')| \\ &\leq \phi \sum_{v,o} Pr(v, o|b, u) \delta_{n-1} \text{ (by def. of } \delta_n) \\ &\leq \phi\delta_{n-1} \sum_{v,o} Pr(v, o|b, u) = \phi\delta_{n-1} \; . \qquad (24) \end{aligned}$$

Second, we prove that $|Q^*_n(b, u) - L_n(b, u)| \leq \epsilon_p |V| R_{\max} \left(1 + 3\phi\frac{|O|}{1-\phi}\right)$:

$$\begin{aligned} |Q^*_n(b, u) - L_n(b, u)| &\leq |R^*(b, u) - R(b, u)| + \\ &\quad \phi \sum_{v,o} |V^*(b^*) Pr^*(v, o|b, u) - V^*(b') Pr(v, o|b, u)| \\ &\leq \epsilon_p |V| R_{\max} + \phi \sum_{v,o} \left(3\epsilon_p \frac{R_{\max}}{1-\phi}\right) \text{ (Lemma 2, Proposition 2)} \\ &= \epsilon_p |V| R_{\max} + 3\epsilon_p |V||O| \frac{\phi R_{\max}}{1-\phi} \\ &= \epsilon_p |V| R_{\max} \left(1 + 3\phi\frac{|O|}{1-\phi}\right) \; . \qquad (25) \end{aligned}$$

Finally, we prove the main result of this proposition:

$$\begin{aligned} |Q^*_n(b, u) - Q_n(b, u)| &\leq |Q^*_n(b, u) - L_n(b, u)| + |L_n(b, u) - Q_n(b, u)| \\ &\leq \phi\delta_{n-1} + \epsilon_p |V| R_{\max} \left(1 + 3\phi\frac{|O|}{1-\phi}\right) \; . \qquad (26) \end{aligned}$$

The last step follows from inequalities (24) and (25) □

Proposition 4. Given our agent's prediction $\pi_{\text{-}t}(s, v) \triangleq Pr(v|s)$ of its counterpart's mixed strategy (i.e., agent -t's mixed strategy) with prediction error ϵ_p, the incurred error δ_n of the corresponding optimal value function V_n, with respect to $Pr(v|s)$, is linearly bounded by ϵ_p:

$$\delta_n \quad \leq \quad \epsilon_p |V| R_{\max} \left[\phi^{n-1} + \frac{1}{1-\phi}\left(1 + \frac{3\phi|O|}{1-\phi}\right)\right] \; .$$

Proof: By Lemma 1, we have:

$$\delta_n \quad \leq \quad \max_b \max_u |Q^*_n(b, u) - Q_n(b, u)| \; . \qquad (27)$$

Also, Proposition 3 shows that:

$$|Q^*_n(b, u) - Q_n(b, u)| \quad \leq \quad \phi\delta_{n-1} + \epsilon_p |V| R_{\max} \left(1 + 3\phi\frac{|O|}{1-\phi}\right) \; . \qquad (28)$$

From (27) and (28), we have:

$$\begin{aligned} \delta_n &\leq \phi\delta_{n-1} + \epsilon_p |V| R_{\max} \left(1 + 3\phi\frac{|O|}{1-\phi}\right) \\ &\leq \phi^{n-1}\delta_1 + \frac{\epsilon_p |V| R_{\max}}{1-\phi} \left(1 + 3\phi\frac{|O|}{1-\phi}\right) \text{ (expanding the recurrence)} \\ &\leq \phi^{n-1}\epsilon_p |V| R_{\max} + \frac{\epsilon_p |V| R_{\max}}{1-\phi} \left(1 + 3\phi\frac{|O|}{1-\phi}\right) \; . \qquad (29) \end{aligned}$$

When $n \to \infty$, $\phi^{n-1}\epsilon_p |V| R_{\max} \to 0$. Thus, with large value of n the incurred error is approximately bounded by $\frac{\epsilon_p |V| R_{\max}}{1-\phi} \left(1 + 3\phi\frac{|O|}{1-\phi}\right)$. Consequently, when $\epsilon_p \to 0$, $\delta_n \to 0$. In general, this implies that the prediction error is linearly proportional with the incurred error of the corresponding value function □

D.3 Main Theorems

Theorem 2. Let V_∞ be the value function for infinite time horizon. Then, we have $\|V_\infty - V_{n+1}\|_\infty \leq \phi\|V_\infty - V_n\|_\infty$.

Proof: We have

$$\begin{aligned}
|V_\infty(b) - V_{n+1}(b)| &\leq \max_u |Q_\infty(b,u) - Q_{n+1}(b,u)| && (30)\\
&\leq \max_u \phi \sum_{v,o} Pr(v,o|b,u)|V_\infty(b') - V_n(b)| && (31)\\
&\leq \max_u \phi \sum_{v,o} Pr(v,o|b,u)\|V_\infty - V_n\|_\infty && (32)\\
&= \phi\|V_\infty - V_n\|_\infty \max_u \sum_{v,o} Pr(v,o|b,u) && (33)\\
&= \phi\|V_\infty - V_n\|_\infty && (34)
\end{aligned}$$

The last equation completes our proof. □

Theorem 3. The optimal value function V_n of I-POMDP Lite is a piecewise-linear and convex function represented as a finite set of α vectors:

$$V_n(b) = \max_{\alpha \in V_n}(\alpha b)\ . \qquad (35)$$

Proof: For $n = 1$, it can be verified that the set of α vectors V_n are simply the weighted average immediate payoffs:

$$V_1 = \left\{\alpha_u \mid u \in U, \forall s \in S, \alpha_u(s) = \sum_v Pr(v|s)R(s,u,v)\right\}\ .$$

Inductively, let us assume that V_n is a piecewise-linear and convex function up to $n = k$. We need to show that V_{k+1} is also a piecewise-linear and convex function, represented as a finite set of α vectors.

To simplify the notation, let us denote $R(s,u) = \sum_v Pr(v|s)R(s,u,v)$ and $b' = B(b,u,v,o)$. Also, let us index the α vectors of V_k with the set of numberings $I = \{1, 2, \ldots, |V_k|\}$. Subsequently, we have

$$V_{k+1}(b) = \max_u\left(\sum_s b(s)R(s,u) + \phi\sum_{v,o} Pr(v,o|b,u)\max_{i\in I}\sum_{s'} b'(s')\alpha'_i(s')\right)\ . \qquad (36)$$

Recall that the belief update step is defined as

$$b'(s') = \frac{1}{Pr(v,o|b,u)}Z(s',u,o)\sum_s T(s,u,v,s')Pr(v|s)b(s)\ .$$

Let l' be the unnormalized version of b', we have

$$l'(s') = Z(s',u,o)\sum_s T(s,u,v,s')Pr(v|s)b(s)\ . \qquad (37)$$

Using equation (37), we can simplify equation (36) as

$$\begin{aligned}
V_{k+1}(b) &= \max_u\left(\sum_s b(s)R(s,u) + \phi\sum_{v,o}\max_{i\in I}\sum_{s'}\alpha'_i(s')l'(s')\right)\\
&= \max_u\left(\sum_s b(s)R(s,u) + \phi\max_{x_{1,1}\in I}\max_{x_{1,2}\in I}\ldots\max_{x_{|V|,|O|}\in I}\sum_{v,o,s'}\alpha'_{x_{v,o}}(s')l'(s')\right)\\
&= \max_u\max_{x_{1,1}\in I}\ldots\max_{x_{|V|,|O|}\in I}\left(\sum_s b(s)R(s,u) + \phi\sum_{v,o,s'}\alpha'_{x_{v,o}}(s')l'(s')\right)\ . && (38)
\end{aligned}$$

Substitute equation (37) into (38), we obtain

$$V_{k+1}(b) = \max_{u} \max_{x_{1,1} \in I} \dots \max_{x_{|V|,|O|} \in I} \left(\sum_{s} b(s) \left(R(s,u) + \phi \sum_{v,o,s'} \alpha'_{x_{v,o}}(s') Z(s',u,o) T(s,u,v,s') Pr(v|s) \right) \right) .$$

The last equation shows that $V_{k+1}(b)$ is a piecewise-linear and convex function. Essentially, it is equivalent to equation (35). □

This proof also implies an exponential increase in the number of α vectors after each back-up operation. Intuitively, for each tuple $(u, x_{1,1}, \dots, x_{|V|,|O|})$, we can compute a new α vector for V_{k+1}. Thus, the number of α vectors needed to exactly represent V_{k+1} is equal to the number of those tuples. Since the domain value for each variable $x_{i,j}$ is the set of integers from $I = \{1, 2, \dots, |V_k|\}$ and there are $|V||O|$ of those variables, we have $|U||V_k|^{|V||O|}$ of such tuples. Consequently, there will be $|V_{k+1}| = |U||V_k|^{|V||O|}$ α vectors generated to represent V_{k+1}. This explains why exact back-up operation with respect to the whole belief simplex is infeasible in practice.

Theorem 4. The performance loss $\delta_n \triangleq \|J^*_n - J_n\|_\infty$ incurred by executing I-POMDP Lite policy, induced with respect to the predicted mixed strategy $\pi_{\text{-}t}(s,v) \triangleq Pr(v|s)$ of agent -t (as compared to its true mixed strategy $\pi^*_{\text{-}t}(s,v) \triangleq Pr^*(v|s)$), after n backup steps is linearly bounded by the prediction error ϵ_p:

$$\|J^*_n - J_n\|_\infty \quad \leq \quad 2\epsilon_p |V| R_{\max} \left[\phi^{n-1} + \frac{1}{1-\phi}\left(1 + \frac{3\phi|O|}{1-\phi}\right)\right] .$$

Proof: Let us define $I_n(b)$ as the expected total reward if our agent follows the optimal policy π computed with respect to our predictive distribution $Pr(v|s)$, and if the other agent's true mixed strategy is exactly $Pr(v|s)$. We have:

$$I_n(b) = \sum_{s,v} Pr(v|s) R(s, \pi(b), v) b(s) + \phi \sum_{v,o} Pr(v, o|b, \pi(b)) I_{n-1}(b') .$$

with $b' = B(b, \pi(b), v, o)$.

It can be trivially verified that $I_n \equiv V_n$. Now, we have

$$\begin{aligned} |J^*_n(b) - J_n(b)| &\leq |J^*_n(b) - I_n(b)| + |I_n(b) - J_n(b)| \\ &= |V^*_n(b) - V_n(b)| + |I_n(b) - J_n(b)| \\ &\leq \epsilon_p C + |I_n(b) - J_n(b)| \text{ (Proposition 4)} . \qquad (39) \end{aligned}$$

with C is a constant that represents for:

$$C \quad = \quad \phi^{n-1}|V|R_{\max} + \frac{R_{\max}|V|}{1-\phi}\left(1 + \frac{3\phi|O|}{1-\phi}\right) . \qquad (40)$$

For convenience, let us denote from now on that $u = \pi(b)$ and $b^* = B^*(b, \pi(b), v, o)$. Thus, we have

$$\begin{aligned} |I_n(b) - J_n(b)| &\leq \sum_{s,v} b(s) R(s,u,v) |Pr^*(v|s) - Pr(v|s)| + \\ &\quad \phi \sum_{v,o} |Pr^*(v,o|b,u) J_{n-1}(b^*) - Pr(v,o|b,u) I_{n-1}(b')| \\ &\leq \phi \sum_{v,o} |Pr^*(v,o|b,u) J_{n-1}(b^*) - Pr(v,o|b,u) I_{n-1}(b')| + \\ &\quad R_{\max}|V|\epsilon_p \text{ (Lemma 2)} \qquad (41) \end{aligned}$$

Next, we have

$$\begin{aligned}
|Pr^*(v,o|b,u)J_{n-1}(b^*) - Pr(v,o|b,u)I_{n-1}(b')| &\leq Pr^*(v,o|b,u)|J_{n-1}(b^*) - I_{n-1}(b')| + \\
&\quad I_{n-1}(b')|Pr^*(v,o|b,u) - Pr(v,o|b,u)| \\
&\leq Pr^*(v,o|b,u)|J_{n-1}(b^*) - I_{n-1}(b')| + \\
&\quad I_{n-1}(b')\epsilon_p \text{ (Lemma 3)} \\
&\leq Pr^*(v,o|b,u)|J_{n-1}(b^*) - I_{n-1}(b')| + \\
&\quad \frac{R_{\max}}{1-\phi}\epsilon_p \qquad (42)
\end{aligned}$$

The above last step follows because the expected total reward $I_{n-1}(b')$ is always bounded by $\frac{R_{\max}}{1-\phi}$. Let us denote $\gamma_n = \max_b |I_n(b) - J_n(b)|$, we have

$$\begin{aligned}
Pr^*(v,o|b,u)|J_{n-1}(b^*) - I_{n-1}(b')| &\leq Pr^*(v,o|b,u)\left(|J_{n-1}(b^*) - I_{n-1}(b^*)| + |I_{n-1}(b^*) - I_{n-1}(b')|\right) \\
&\leq Pr^*(v,o|b,u)\left(|J_{n-1}(b^*) - I_{n-1}(b^*)| + |V_{n-1}(b^*) - V_{n-1}(b')|\right) \\
&\leq Pr^*(v,o|b,u)\left(|J_{n-1}(b^*) - I_{n-1}(b^*)| + \frac{R_{\max}}{1-\phi}\|b^* - b'\|\right) \text{ (Lemma 4)} \\
&\leq Pr^*(v,o|b,u)\left(\gamma_{n-1} + \frac{R_{\max}}{1-\phi}\|b^* - b'\|\right) \text{ (Def. of } \gamma_n\text{)} \\
&\leq Pr^*(v,o|b,u)\gamma_{n-1} + \frac{R_{\max}}{1-\phi}Pr^*(v,o|b,u)\|b^* - b'\| \\
&\leq Pr^*(v,o|b,u)\gamma_{n-1} + \frac{R_{\max}}{1-\phi}2\epsilon_p \text{ (Proposition 1)} \\
&= Pr^*(v,o|b,u)\gamma_{n-1} + 2\epsilon_p\frac{R_{\max}}{1-\phi} \qquad (43)
\end{aligned}$$

Substitute (43) into (42), we have

$$|Pr^*(v,o|b,u)J_{n-1}(b^*) - Pr(v,o|b,u)I_{n-1}(b')| \leq Pr^*(v,o|b,u)\gamma_{n-1} + 3\epsilon_p\frac{R_{\max}}{1-\phi}\ . \qquad (44)$$

Substitute inequality (44) into (41), we have

$$\begin{aligned}
|I_n(b) - J_n(b)| &\leq R_{\max}|V|\epsilon_p + \phi\sum_{v,o}\left(Pr^*(v,o|b,u)\gamma_{n-1} + 3\epsilon_p\frac{R_{\max}}{1-\phi}\right) \\
&\leq R_{\max}|V|\epsilon_p + \phi\sum_{v,o}Pr^*(v,o|b,u)\gamma_{n-1} + 3\phi|V||O|\frac{R_{\max}}{1-\phi}\epsilon_p \\
&\leq \phi\gamma_{n-1} + R_{\max}|V|\epsilon_p + 3\phi|V||O|\frac{R_{\max}}{1-\phi}\epsilon_p\ . \qquad (45)
\end{aligned}$$

Since inequality (45) holds for all b, we have the recurrence equation

$$\begin{aligned}
\gamma_n &\leq \phi\gamma_{n-1} + \epsilon_p|V|R_{\max}\left(1 + \frac{3\phi|O|}{1-\phi}\right) \\
&\leq \phi^{n-1}\gamma_1 + \frac{\epsilon_p|V|R_{\max}}{1-\phi}\left(1 + \frac{3\phi|O|}{1-\phi}\right) \text{ (expand the recurrence)} \\
&\leq \phi^{n-1}\epsilon_p|V|R_{\max} + \frac{\epsilon_p|V|R_{\max}}{1-\phi}\left(1 + \frac{3\phi|O|}{1-\phi}\right) \\
&= \epsilon_p\left(\phi^{n-1}|V|R_{\max} + \frac{R_{\max}|V|}{1-\phi}\left(1 + \frac{3\phi|O|}{1-\phi}\right)\right) \\
&= \epsilon_p C\ . \qquad (46)
\end{aligned}$$

From (46), it is obvious that $|I_n(b) - J_n(b)| \leq \gamma_n \leq \epsilon_p C$. Hence, plugging it into inequality (39), we have

$$|J^*_n(b) - J_n(b)| \leq \epsilon_p C + |I_n(b) - J_n(b)| \leq 2\epsilon_p C\ . \qquad (47)$$

Since inequality (47) holds for all b, we have:

$$\|J_n^* - J_n\|_\infty = \max_b |J_n^*(b) - J_n(b)| \leq 2\epsilon_p C \ . \tag{48}$$

Finally, we complete the proof by substituting equation (40) into inequality (48):

$$\|J_n^* - J_n\|_\infty \quad \leq \quad 2\left(\phi^{n-1}\epsilon_p|V|R_{\max} + \frac{\epsilon_p|V|R_{\max}}{1-\phi}\left(1 + 3\phi\frac{|O|}{1-\phi}\right)\right) \ . \tag{49}$$

□